Temporal Anomaly Detection in Multi-Spectral Satellite Imagery

*MSDS Capstone Project - April 2025

https://github.com/Shakeri-Lab/ANOMALOUS.git

Kiana Dane
University of Virginia
urnh8e@virginia.edu

Devon Hathaway
University of Virginia
wtc6vz@virginia.edu

Zachary Stautzenbach
University of Virginia
ehe5bn@virginia.edu

Jonathan Swap
University of Virginia
js7jx@virginia.edu

I. ABSTRACT

This project advances the field of anomaly detection in remote sensing by analyzing temporal patterns in multispectral satellite imagery. We explored three main approaches: Functional Principal Component Analysis (fPCA), the Prithvi-EO-2.0 foundational model (Prithvi), and Continuous Change Detection and Classification (CCDC). Our main focus was on the fPCA model, which achieved 66% accuracy and an F1 score of 0.30 for anomaly detection.

Despite challenges—including the need for temporal downsampling, artifacts in the Planet.com imagery (such as sensor-induced banding or haze not attributable to clouds), and limited ground truth labels (only 3–5% manually annotated)—the fPCA model demonstrated strong performance, particularly for sparse time series data. It is computationally efficient, requiring only 3–4 minutes per site, making it well-suited for environments where data are regularly updated and the model must be re-run frequently.

Although Prithvi achieved slightly higher accuracy, its precision, recall, and F1 scores were lower, and its run times were highly variable. Nonetheless, with further development and optimization, Prithvi has the potential to become a competitive alternative. CCDC performed moderately well, with stable accuracy across sites, but its sensitivity to input parameters and reliance on having only a few temporal gaps limited its performance in noisier datasets.

All three models—fPCA, Prithvi, and CCDC—would benefit from consistent preprocessing and fully harmonized datasets to enable more direct and fair comparisons. Future work will focus primarily on enhancing the fPCA and Prithvi models, with an emphasis on improving noise reduction techniques and developing more robust anomaly detection algorithms.

II. INTRODUCTION

Monitoring the Earth's surface over time is essential for various applications, such as detecting deforestation, identifying crop phenology changes, tracking urban expansion, and assessing the impacts of climate change. Temporal anomalies, which represent unexpected changes in observed spectral signals, can indicate significant environmental events, ecosystem disturbances, or anthropogenic changes. Despite recent advances, there is a need for an open-source, scalable tool that combines the precision of classical time-series algorithms with the adaptability of deep learning.

We explored several methods for detecting anomalies. We leveraged dimensionality reductions, pattern recognition algorithms, and deep learning approaches against current industry standard methodology. This paper seeks to address each of our approaches and evaluation metrics.

III. RELATED WORK

There has been extensive work to create algorithms to detect anomalies in satellite imagery. Existing algorithms include LandTrendr, Breaks For Additive Season and Trend (BFAST), and CCDC. Their primary purpose is to detect land changes over time while accounting for seasonal variations. Recent studies have been done to understand how to apply fPCA to detect variation in land as well. For example, Pesaresi et al. (2020) leveraged the time-series capability of fPCA to understand plant associations in a forest in Marche, Italy [1]. Modeling Normalized Difference Vegetation Index (NDVI) curves, fPCA summarized the primary temporal pattern and quantified variation through component scores. These scores flag unusual behavior that occurs over time. Although their study primarily focused on habitat classification rather than anomaly detection, their temporal modeling approach aligns closely with our objective. In an additional paper, Peraresi et al. applied a similar fPCA method to Mount Canero to capture continuous seasonal vegetation patterns further reinforcing the capability of fPCA in identifying deviations from seasonal behavior [2].

Another approach our team identified was to adapt the methodology introduced by Szwarcman et al.[3], a collaboration between IBM and NASA that resulted in the development of Prithvi-EO-2.0, a pretrained foundational model based on a Vision Transformer (ViT) architecture. This model was trained on a vast dataset comprised of 4.2 million global time series samples from NASA's Harmonized Landsat and

Sentinel-2 (HLS) archives at 30m² spatial resolution. With approximately 300 million trainable parameters, Prithvi-EO-2.0 has demonstrated strong performance across a range of Earth Observation tasks. The use case presented in their paper that most closely aligns with our objectives is the multi-temporal crop classification task, where the model was trained and evaluated across 13 crop classes over 80 epochs, achieving an F1 score of 84.4% using the full training dataset. Given the model's demonstrated capacity to learn spatiotemporal representations, the Prithvi model is a good candidate for adaptation to anomaly detection in hyperspectral satellite imagery.

IV. Data Description

The data used in this project was acquired from planet.com, a satellite image repository that provides researchers with up-to-date high-resolution satellite imagery from around the world. 14 test sites from around the world were selected. The diversity of latitudes of these locations introduces more seasonal variation, encouraging our models to understand general patterns of anomalies rather than the spectral values of the specific pixels in those individual locations. Some of these sites were chosen for known anomalies, while others served as control sites (Table II). Our dataset consisted of over 2300 images across these 14 sites.

The images were captured by SuperDove satellites and have eight bands of data, meaning that light reflectance values were recorded for each pixel for eight different wavelengths of light (coastal blue, blue, green I, green II, yellow, red, red-edge, and near-infrared). These bands are comparable to those found in well-known Earth observation satellites such as Landsat and Sentinel-2, which also capture multispectral imagery. See Table I for a comparison of features across these three types of satellites. SuperDove satellites were first launched in 2020, meaning that all data collected for this experiment ranges in date from 2020 to 2025.

Satellite	No. of Bands	Spatial Resolution	Image Frequency
SuperDove	8	3 m	Near-daily
Sentinel-2	13	10–60 m	5 days
Landsat	11	30 m	16 days

TABLE I: Comparison of SuperDove, Sentinel-2, and Landsat imagery characteristics

Each image from Planet.com included an 8-banded TIFF image of the site, as well as a smaller TIFF that acted as a mask. The bands of this mask included metadata such as cloud cover, snow cover, and any pixels that were not viable. This metadata allowed us to remove poor quality images from the dataset.

Our goal was to acquire at least one image per month for each site over the course of the five years of data we had access to. Team members would manually choose images to include in the dataset and filter out images with 25% cloud cover, or images acquired by any satellite other than SuperDove.

V. Preprocessing

We preprocessed our data from each site of interest and down sampled the spatial resolution from 3 m^2 to 30 m^2 per pixel. Down sampling was necessary to reduce the strain on our computing resources, and make labeling the data feasible. The preprocessing pipeline consisted of the following steps:

1) Extracted images from Rivanna storage by filtering for location and date, yielding:
   - Primary TIFFs with spectral band reflectance
   - Cloud cover mask TIFFs
   - Metadata including capture dates
2) Converted TIFFs into 3D NumPy arrays (height × width × bands).
3) Downsampled image data using `block_reduce()` from the `scikit-image` package with mean aggregation. A quality mask was also generated to track data reliability.
4) Cropped images to remove null or missing reflectance values.
5) Verified alignment of file names with correct image capture dates.
6) Transformed the resulting 4D tensors (time × height × width × bands) into 2D arrays suitable for model input.
7) Compressed and saved the processed data as `.csv.gz` files for efficient storage and future use.

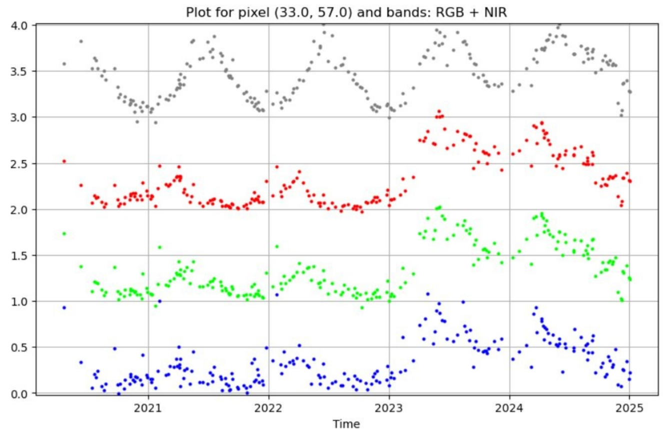


Fig. 1: Plot of a single anomalous pixel over time. Anomaly occurs in early 2023.

A. Anomaly Labeling

Our team did not have access to pre-labeled images and were therefore responsible for creating our own training data. We employed a mixture of labeling by hand and automatically. We plotted pixel reflectance values for red, green, blue and near infrared over time for a single pixel and determined if it was anomalous or normal depending on the seasonal changes throughout the five years of available data (Figure 1). Labeling was based on these four bands for increased human interpretability and because RGB + NIR were the only bands compatible with the Prithvi model.

We repeated this process of labeling pixels by hand until approximately 3-5% of all pixels had been labeled. We then automatically classified the rest by giving pixels the same

Site Name	Location	No. of Images	Size (km^2)	Justification
Apui	Brazil	81	17.86	Rainforest Deforestation
Armazones	Chile	247	4.99	Construction of Observatory
Black Forest	Germany	308	15.75	Beetle Causing Mass Tree Die-Off
Charlottesville	Virginia, USA	242	7.76	Construction of Subdivision
Choros	Chile	103	14.59	Semiarid, Control
Council Bluffs	Iowa, USA	179	10.42	Cropland, Control
Foresthill	California, USA	233	9.52	Forest Fire
Garren Creek	North Carolina, USA	113	13.20	Landslides
Kakamega	Kenya	181	18.76	Rainforest, Control
La Palma	Spain	67	94.82	Volcano Eruption
Limpopo	South Africa	187	10.11	Development
Pokrovsk	Ukraine	193	19.17	Warzone
Surf	Florida, USA	79	18.78	Coastal Region, Control
Upsala	Chile	88	16.42	Melting Glacier

TABLE II: Site information including location, size, and justification for selection.

label as the already labeled pixel with the shortest Euclidean distance and repeated this for all sites in the dataset. This resulted in highly accurate training and testing data. The output of this work was two TIFF images: one 8-banded image of the site, and a single banded mask that included the labeled pixels for the site.

VI. Methodology

We explored three modeling approaches, fPCA, a deep learning approach using the Prithvi-EO-2.0 foundation model, and CCDC.

A. Functional Principal Component Analysis

We looked to apply the algorithm fPCA, an extension of Principal Component Analysis (PCA) by taking into account time-series data even if our observations for a site are spread out unevenly across the years. This algorithm is very advantageous when dealing with a large dataset. It helps simplify complex time-series data by identifying a few main patterns, called functional principal components, that describe most of the variation. Unlike PCA where components represent vectors, components of fPCA are smoothing functions or curves that represent the main directions of variation in the data in a continuous space. It maps the data onto smoothing curves that can capture how major trends and patterns evolve continuously over time. The output is a set of scores for each observation that show how well it aligns with the main pattern for that specific component. In our case, each pixel in the satellite images was assigned a score. If the absolute value of a score was extreme, it flagged that pixel as potentially anomalous, meaning it could represent unusual behavior. With respect to our data, we were able to run an fPCA model utilizing all eight bands, the corresponding pixel coordinates, and the time stamp. On top of all of its analytical capabilities, it is very computationally efficient and its smoothing helps denoise the data such as reducing the impact of transient artifacts such as cloud. The workflow of this process is explained in the following sections.

1) Step 1: Initial fPCA Application and Observation

The first step in applying this algorithm was inputting all the eight-banded time-stamped data for each location into an fPCA model. We used four total components to represent the data. This approach gave us scores for each observation of each component per year. Scores that were very high or very low (negative) highlighted possible anomalies, and scores closer to 0 were typically less anomalous. Plots to view the scores spatially were promising in that it did highlight areas of known anomalies. The early years did not show much since the known anomaly did not happen while the later years highlighted many anomalies in the plot. However, this initial approach led to a lot of noise, which results in a lot of false positives.

2) Step 2: Standard Deviation of the Scores

The next step was to calculate the standard deviation of each pixel over the five year time period. When we group by pixel and take the standard deviation column-wise, we can see how each pixel's behavior varied across the whole time period for each FPC within their respective component. Scores with a higher standard deviation differ more from the average, potentially flagging more anomalous behavior. Once again, these standard deviations of the scores were spatially plotted allowing for us to visualize which parts of the image experienced the most temporal variability, or fluctuation over time.

3) Step 3: Applying Euclidean Norm

In order to use a more explainable and summarized metric for total variability that we explored with the FPCs, we calculated the Euclidean Norm, as shown below in Equation 1 where n is the n^{th} component.

$$\text{Euclidean Norm} = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2} \tag{1}$$

$$\text{Threshold} = b \cdot \text{Quantile}_a(\text{Euclidean Norm Score}) \tag{2}$$

The Euclidean Norm outputs a single value for each pixel. This allows us to use different thresholding metrics to decide which pixels to label anomalous. We developed an equation to threshold anomalies depending on the anomaly prevalence rate, as shown above in Equation 2 where a is typically the mean or median of the Euclidean Norm scores while b is a tuning parameter to adjust the sensitivity for flagging pixels as anomalies. The equation assumes some pixels are normal and looks for higher numbers in the distribution of the Euclidean Norm values.

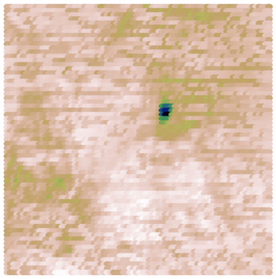


Fig. 2: Euclidean Norm Scores Visualized

We leveraged the ground truth labels to evaluate how well our model and the Euclidean Norm thresholding identified anomalous pixels. We calculated our evaluation metrics for each of our individual sites and then averaged them to see which a and b values provided optimal metrics. If there are no ground truth labels to help grid searching for optimal parameters, then one will have to plot and qualitatively adjust for optimal results. Figure 2 shows the Euclidean Norm scores spatially for one of the area of interest, Armazones, Chile, which has been the construction site of a large astronomical facility since 2014. The darker colors, blue and green, are higher scores possibly flagging an anomaly. The blue spot is where the facility has been built.

B. Prithvi Foundational Model

To ensure compatibility with the Prithvi-EO-2.0 model, both the data and the model were reshaped so that only four bands (red, blue, green, and near infrared) were included. Prithvi expects square input patches of a fixed size. To satisfy this requirement, we divided each satellite image and its corresponding labeled mask into 56×56 pixel tiles.

To use the model for temporal analysis, we prepared the data by stacking multiple timestamps into a single input array. Images acquired during the same month across five consecutive years were stacked, yielding inputs with five time steps. To evaluate model generalization, we divided the dataset into four distinct train/test splits, ensuring that each site appeared exactly once in the test set. Strict data separation was maintained to prevent data leakage between training and evaluation.

The Prithvi model is pretrained using a masked autoencoder (MAE) strategy. The algorithm introduces 3D patch embeddings and 3D positional encodings that account for the spatial (height and width) and temporal (time) dimensions of satellite image sequences. Additionally, the model incorporates auxiliary metadata (latitude and longitude) through independent 2D sinusoidal embeddings, which are added as biases to the embedded tokens. To improve robustness, metadata dropout is applied during pretraining, randomly omitting temporal or location information.

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|M|} \sum_{i \in M} \|x_i - \hat{x}_i\|^2 \tag{3}$$

During fine-tuning, we employed a weighted cross-entropy loss function, with class weights inversely proportional to the anomaly rate in the training set (excluding the validation and testing data). An AdamW optimizer was used for optimization. Random horizontal and vertical flips were applied as data augmentation strategies during training. Due to the lack of early stopping mechanisms in the current Prithvi fine-tuning framework, we manually experimented with different epoch counts. Based on validation performance, eight training epochs were selected as the optimal point to balance learning progress and overfitting.

C. Continuous Change Detection and Classification

Our third approach applies CCDC, a widely adopted baseline in the remote sensing community[4]. CCDC continuously monitors land surface dynamics using harmonic regression, enabling consistent detection of both gradual and abrupt changes at high temporal resolution.

By aligning with CCDC's performance benchmarks, we set a practical goal of achieving comparable speed and accuracy to the original algorithm introduced by Zhu and Woodcock (2014), while incorporating select updates inspired by the COLD framework, such as the use of higher-order harmonic terms to better fit complex seasonal signals.

Harmonic Models for Seasonal Signal Fitting

The input variable x denotes time in fractional years. We used our custom CCDC algorithm to fit a third-order harmonic regression model to the historical time series of each pixel. The residuals from this fit are monitored over time: when the observed data deviates significantly from the fitted model, a land surface change (anomaly) is flagged.

$$\begin{aligned} f(x) = a + bx &+ c \cdot \cos(2\pi x) + d \cdot \sin(2\pi x) \\ &+ e \cdot \cos(4\pi x) + f \cdot \sin(4\pi x) \\ &+ g \cdot \cos(6\pi x) + h \cdot \sin(6\pi x) \end{aligned} \tag{4}$$

This third-order harmonic model captures seasonal variation with increasing complexity by including up to three harmonic terms. Simpler models, such as the first- and second-order harmonics, are nested within this formulation: the first-order model includes only the annual cycle $(\cos(2\pi x), \sin(2\pi x))$, while the second-order model adds the semi-annual cycle $(\cos(4\pi x), \sin(4\pi x))$.

By truncating the equation at different harmonic levels, we control the model's ability to represent fine-grained seasonal dynamics, which is especially important when data quality or availability is limited (e.g., due to cloud cover or shorter time series), higher harmonic orders can lead to overfitting or unstable fits. In such cases, reducing the harmonic level (e.g., using `har1` or `har2`) provides a more stable seasonal representation and often yields better performance.

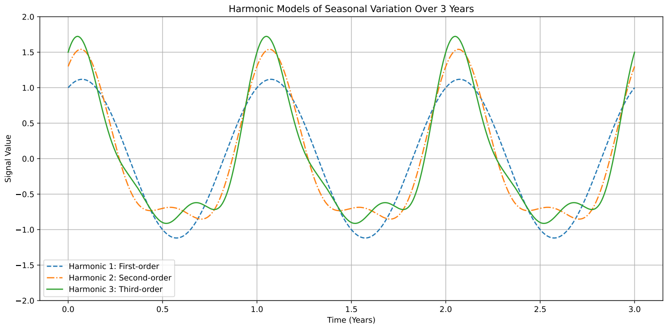


Fig. 3: Harmonic models of increasing complexity over three years

VII. Results

When evaluating model performance, we aimed to assess not only the overall classification accuracy but also the model's ability to correctly detect rare anomalies, which is critical in anomaly detection tasks.

Thus, in addition to accuracy, we report precision, recall, F1 score, and run-time. Anomalies in the wild tend to be rare, so these metrics provide a more nuanced view of performance. Traditional accuracy alone may be misleading if a model is classifying everything as the dominant class. In particular, precision and recall are important indicators of how well the model identifies true anomalies while avoiding false detections.

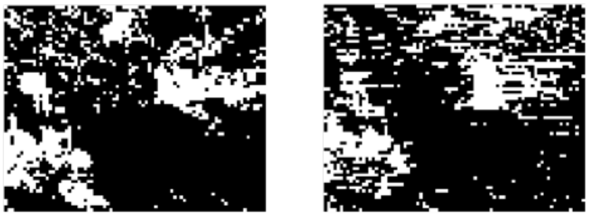
Table III compares these evaluation metrics among the three models.

Metric	**Prithvi (%)**	**fPCA (%)**	**CCDC (%)**
Overall Accuracy	73.1	66.0	79.4
Precision	26.4	31.0	43.1
Recall	32.9	59.0	31.8
F1 Score	21.9	30.0	29.6
Run-Time	10-360/site min	3-5 min/site	15-45 min/site

TABLE III: Comparison of average reporting metrics across models

All of the models achieved relatively decent accuracy but compared F1 score to understand the balance between precision and recall to capture both the model's ability to detect relevant instances and avoid misclassifications. Through benchmarking, fPCA shows the most promise and success for detecting anomalies due to the previously mentioned metrics, and its efficient run-time of three to five minutes per site.

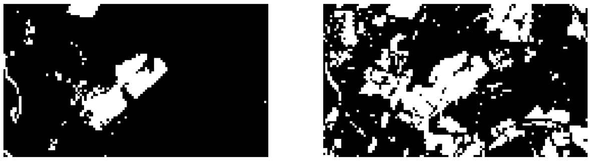
Figure 4 shows the ground truth vs. the fPCA predicted labels of anomalies in Armazones, Chile that was mentioned previously. Figure 5 shows the same information but for North Pointe, a new housing development in Charlottesville, VA. The algorithm effectively captures many anomalies, though it also picks up some noise especially in areas with rarer surface types such as urban areas and meadows.



(a) Ground Truth Labels

(b) Predicted Labels

Fig. 4: Armazones, Chile Ground Truth vs. Predicted, White = anomalous; black = normal.



(a) Ground Truth Labels

(b) Predicted Labels

Fig. 5: North Pointe, Charlottesville Ground Truth vs. Predicted, White = anomalous; black = normal.

VIII. Discussion

A few key limitations in this project are the striping from Planet.com and ground truth labels, the model requiring down sampled data, and the generalization of the model.

Planet.com has some striping issues, which we could see in our images, causing noise and lower evaluation metrics. This coincides with the ground truth labels we had, which were partially hand-labeled and then automated. The fact that not all of the ground truth labels were hand-labeled means there is some potential error that could mislabel pixels that should be labeled anomalous but were not. This was another area that may have lowered our evaluation metrics.

Additionally, another key limitation is the necessity of down sampling our data. We were unable to utilize the finer grain resolution because the model itself could not handle the sheer magnitude of full eight-band image data. Any addition of new data would also have to be down sampled and recompiled, requiring the model to be rerun.

Finally, fPCA was limited by its generalizability. As mentioned earlier in the paper, anomaly detection has no specific or constant anomaly ratio for different sites. Each site is specific to its own context, which we saw in the distribution of our own sites, so our model will have different thresholding parameters for each site you run it on.

IX. Conclusion and Future Work

Overall, of our three approaches, our strongest model is the fPCA. Despite its slightly lower overall accuracy, it has a higher F1 score, which balances the recall and precision. Additionally, the recall and precision metrics are both higher in the fPCA model as opposed to Prithvi model. This means that the fPCA model overall can more effectively catch the anomalous pixels, which is more important than just the raw accuracy.

While this study has demonstrated the feasibility of using machine learning to detect temporal anomalies in multispectral satellite imagery, several important areas remain for future development and refinement.

A key next step is to improve how thresholds are defined for flagging anomalous behavior. Current thresholds may under- or over-identify significant events depending on the landscape or seasonal variability. Future iterations of the model should incorporate adaptive thresholding techniques that account for local temporal dynamics, environmental variability, and land cover-specific baselines. Incorporating statistical control methods or learning-based threshold optimization could help balance sensitivity and specificity more effectively across diverse regions and time periods.

Currently, the model flags anomalous activity without differentiating between causes or categories. Introducing a secondary classification step would allow the system to distinguish between anomaly types—for instance, separating natural events (e.g., floods, wildfires) from anthropogenic changes (e.g., construction, deforestation). This classification capability would add critical context to anomaly alerts, increasing their relevance and actionability for end users.

Our current analysis operates at the pixel level, which limits spatial context and interpretability. Future work will explore scaling the approach to object-level anomaly detection, enabling the identification, segmentation, and temporal tracking of meaningful geospatial features.

The data preprocessing workflow—including cloud masking, image harmonization, and manual quality assurance—remains time-consuming and labor-intensive. Future work should prioritize the automation of these tasks to increase throughput and reproducibility. Developing rule-based or machine learning Quality Assurance tools, version-controlled data ingestion pipelines, and standardized metadata tagging systems will enable more scalable and transparent workflows.

For anomaly detection to support operational monitoring—such as in disaster response or environmental compliance—models must be capable of processing and flagging changes in near-real-time. Achieving this will require optimizing the end-to-end pipeline for latency and integrating data streams that can be updated continuously.

Finally, the integration of ancillary datasets could enhance both anomaly detection accuracy and interpretability. Incorporating weather data, topographic variables, human activity layers (e.g., population density or road networks), and ecological indicators could help the model distinguish between normal variation and meaningful change. Multi-source data fusion would also open up opportunities for context-aware modeling, particularly in heterogeneous or dynamic environments.

In sum, these future directions offer clear pathways to strengthen the model's accuracy, scalability, operational relevance, and scientific value. Addressing them will further advance the goal of detecting meaningful geospatial changes with precision, speed, and contextual awareness.

Acknowledgments

We would like to acknowledge the long hours, work, and mentorship that our mentor Dr. Tobias Fritz has invested in supporting this project. Additionally, thank you to our sponsor Dr. Heman Shakeri for his guidance and feedback.

References

[1] Simone Pesaresi, Adriano Mancini, Giacomo Quattrini, and Simona Casavecchia. Mapping mediterranean forest plant associations and habitats with functional principal component analysis using landsat 8 ndvi time series. *Remote Sensing*, 12(7):1132, 2020. ISSN 2072-4292. doi: 10.3390/rs12071132. URL `https://www.mdpi.com/2072-4292/12/7/1132`.

[2] Simone Pesaresi, Adriano Mancini, Giacomo Quattrini, and Simona Casavecchia. Evaluation and selection of multi-spectral indices to classify vegetation using multivariate functional principal component analysis. *Remote Sensing*, 16(7), 2024. ISSN 2072-4292. doi: 10.3390/rs16071224. URL `https://www.mdpi.com/2072-4292/16/7/1224`.

[3] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Carlos Gomes, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Disha Shidham, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, David Bell, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications, 2025. URL `https://arxiv.org/abs/2412.02732`.

[4] Zhe Zhu, Junxue Zhang, Zhiqiang Yang, Amal H. Aljaddani, Warren B. Cohen, Shi Qiu, and Congliang Zhou. Continuous monitoring of land disturbance based on landsat time series. *Remote Sensing of Environment*, 238:111116, 2020. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2019.03.009. URL `https://www.sciencedirect.com/science/article/pi`
Time Series Analysis with High Spatial Resolution Imagery.